PATENT ABSTRACTS OF JAPAN

(11)Publication number : **11-085615**
(43)Date of publication of application : **30.03.1999**

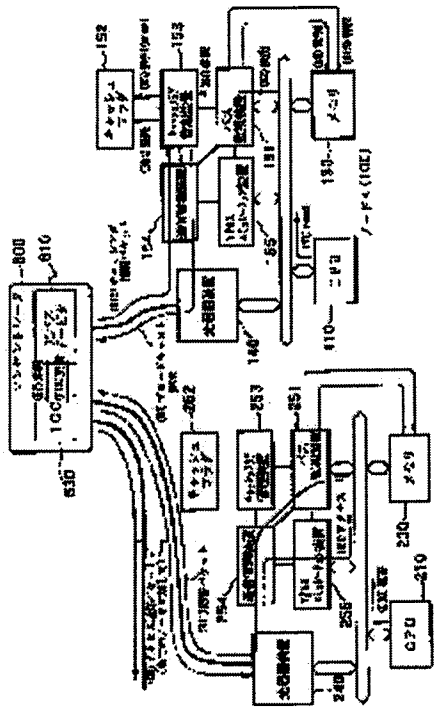
(51)Int.Cl. G06F 12/08
G06F 15/163
G06F 15/173

(21)Application number : **09-243127**
(22)Date of filing : **08.09.1997**

(71)Applicant : **CANON INC**
(72)Inventor : **SHIMOYAMA TOMOHIKO**

(54) SYSTEM AND DEVICE FOR PROCESSING INFORMATION AND CONTROL METHOD THEREFOR

(57)Abstract:
PROBLEM TO BE SOLVED: To reduce the capacity of a storage mechanism for recording the destination to multicast access information required for providing a distributed shared memory.
SOLUTION: Nodes A and B for caching data provided by access to a shared memory space hold cache flags 152 and 252 showing whether their own memory spaces are cached in the other nodes or not for the unit of a cache. When the access to a memory in the present node is detected, the cache flag is referred to and when the relevant access address is cached by the other node, the multicast of the relevant access is requested to a concentrator 600. An ICC 630 of the concentrator 600 holds the cache of more detailed directory information and in the case of cache hit, the relevant access is reported to the node shown by this directory information.



D1

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－85615

(43)公開日　平成11年(1999)３月30日

(51)Int.Cl.[6]		識別記号	ＦＩ		
Ｇ０６Ｆ	12/08	３１０	Ｇ０６Ｆ	12/08	３１０Ｂ
	15/163			15/16	３２０Ｋ
	15/173				４００Ｎ

審査請求　未請求　請求項の数14　ＯＬ　（全 17 頁）

(21)出願番号　特願平9－243127

(22)出願日　平成９年(1997)９月８日

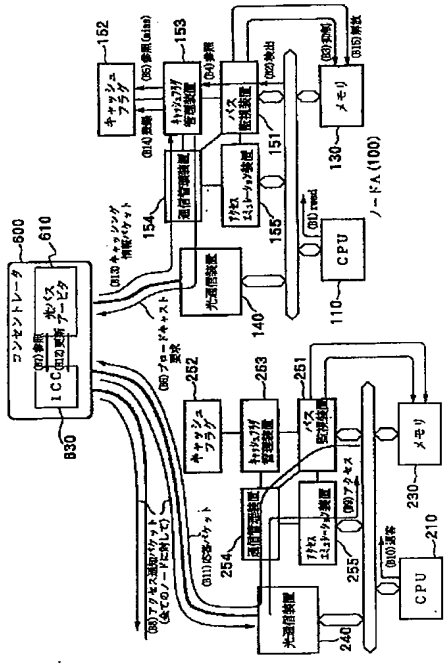
(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号

(72)発明者　下山　朋彦
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(74)代理人　弁理士　大塚　康徳　（外２名）

(54)【発明の名称】　情報処理システム及び情報処理装置及びその制御方法

(57)【要約】
【課題】分散共有メモリを実現する際に必要となるアクセス情報のマルチキャストの宛て先を記録するための記憶機構の容量を削減する。
【解決手段】共有メモリ空間へのアクセスによって得られたデータをキャッシングするノードＡ、Ｂは、自身のメモリ空間に関して他のノードにキャッシングされているか否かをキャッシュ単位で示すキャッシュフラグ１５２、２５２が保持されている。自ノード内のメモリに対するアクセスが検出されるとキャッシュフラグを参照し、当該アクセスアドレスが他ノードにキャッシングされている場合は、コンセントレータ６００に当該アクセスのマルチキャストを要求する。コンセントレータ６００のＩＣＣ６３０にはより詳細なディレクトリ情報のキャッシュが保持されており、キャッシュヒットした場合は、このディレクトリ情報で示されるノードへ当該アクセスが通知される。



1

【特許請求の範囲】

【請求項１】　複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムであって、

前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ手段と、

前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、

前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段とを備えることを特徴とする情報処理システム。

【請求項２】　前記メンテナンス手段は、

前記複数の情報処理装置の各々において、自身のメモリ空間へのアクセスが発生した場合に、当該アクセス先アドレスが他の情報処理装置にキャッシングされているか否かを前記キャッシュ情報を参照して判定する判定手段と、

前記判定手段によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定された場合、当該アクセスをマルチキャストするマルチキャスト手段とを備えることを特徴とする請求項１に記載の情報処理システム。

【請求項３】　前記複数の情報処理装置の一つが、キャッシングアドレスとキャッシング先の情報処理装置を示すマルチキャスト情報を保持する第２保持手段を備える通信管理装置であり、

前記マルチキャスト手段による前記アクセスのマルチキャストは、前記第２保持手段に保持されたマルチキャスト情報に基づいて選択された情報処理装置に対して行なわれることを特徴とする請求項２に記載の情報処理システム。

【請求項４】　前記マルチキャスト手段は、前記判定手段によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定され、前記第２保持手段に対応する情報が保持されていない場合、当該アクセスを前記複数の情報処理装置の全てにマルチキャストすることを特徴とする請求項３に記載の情報処理システム。

【請求項５】　前記複数の情報処理装置を接続する通信網が、前記通信管理装置を中心としたスター結合であることを特徴とする請求項３に記載の情報処理システム。

【請求項６】　前記複数の情報処理装置を接続する通信網が、複数の波長の光を用いて接続する光波長多重化した経路により構成されることを特徴とする請求項１乃至５のいずれかに記載の情報処理システム。

【請求項７】　複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置であって、

前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ手段と、

自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、

前記メモリ空間にアクセスが発生した場合、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段とを備えることを特徴とする情報処理装置。

【請求項８】　前記メンテナンス手段は、

前記複数の情報処理装置の各々において、自身のメモリ空間へのアクセスが発生した場合に、当該アクセス先アドレスが他の情報処理装置にキャッシングされているか否かを前記キャッシュ情報を参照して判定する判定手段と、

前記判定手段によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定された場合、当該アクセスのマルチキャストを要求する要求手段とを備えることを特徴とする請求項７に記載の情報処理装置。

【請求項９】　複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムの制御方法であって、

前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ工程と、

前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、

前記保持工程によって保持されたキャッシュ情報に基づいて前記キャッシュ工程のキャッシュメンテナンスを行うメンテナンス工程とを備えることを特徴とする制御方法。

【請求項１０】　前記メンテナンス工程は、

前記複数の情報処理装置の各々において、自身のメモリ空間へのアクセスが発生した場合に、当該アクセス先アドレスが他の情報処理装置にキャッシングされているか否かを前記キャッシュ情報を参照して判定する判定工程と、

前記判定工程によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定された場合、当該アクセスをマルチキャストするマルチキャスト工程とを備えることを特徴とする請求項９に記載の制御方法。

【請求項１１】　前記マルチキャスト工程は、前記判定工程によって前記アクセス先アドレスが他の情報処理装

置によってキャッシングされていると判定された場合、当該アクセスを前記複数の情報処理装置の全てにマルチキャストすることを特徴とする請求項９に記載の制御方法。

【請求項１２】　前記複数の情報処理装置の一つが、キャッシングアドレスとキャッシング先の情報処理装置を示すマルチキャスト情報をメモリに保持する第２保持工程を備える通信管理装置であり、

前記マルチキャスト工程による前記アクセスのマルチキャストは、前記第２保持工程によって前記メモリ保持されたマルチキャスト情報に基づいて選択された情報処理装置に対して行なわれることを特徴とする請求項１１に記載の制御方法。

【請求項１３】　複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置の制御方法であって、

前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ工程と、

自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、

前記メモリ空間にアクセスが発生した場合、前記保持工程によって保持されているキャッシュ情報に基づいて前記キャッシュ工程におけるキャッシュメンテナンスを行うメンテナンス工程とを備えることを特徴とする制御方法。

【請求項１４】　複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムに適用可能な情報処理装置のための制御プログラムを格納する記憶媒体であって、該制御プログラムがコンピュータを、

前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ手段と、

自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、

前記メモリ空間にアクセスが発生した場合、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段として機能させることを特徴とする記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は分散共有メモリとその上でのキャッシュ機構を有する情報処理システム及び情報処理装置及びその制御方法に関するものである。

【０００２】

【従来の技術】分散共有メモリとその上でのキャッシュ機構を有する情報処理システムが提案されている。一般に、分散共有メモリを実現する場合、共有メモリの一貫

性保持のために必要な通信を必要な計算機にのみ配送するため、あるアクセスに関しての情報をどの計算機に送るかを記録する必要があった。例えばディレクトリ方式の分散共有メモリのキャッシュシステムにおいては、メモリのキャッシュライン毎にどの計算機にアクセス情報をマルチキャストしなければならないかを記録している。

【０００３】だがどの計算機にマルチキャストするかを示す宛て先の保持に使用するメモリの容量は、通信機構上、計算機の数により増大してゆき、その容量を押さえることが求められている。一般に、容量を押さえるための手法として、宛て先の情報を欠落させ、不必要な計算機にまでマルチキャストを行うことが行われることが多かった。しかしながら、この手法では、記憶容量を節約できる反面、不必要な計算機にマルチキャストを行うことによる性能の低下が避けられなかった。

【０００４】これを補う技術として、頻繁に使用されるメモリアドレスについては詳しく宛て先を記録し、他のメモリアドレスに関しては宛て先の情報を欠落させる（特願平８－１１９２６８）、頻繁に使用しされるメモリアドレスについては宛先を記録し他のメモリアドレスに関してはブロードキャストを行う（特願平８－３１０２２）、といった手法が本出願人によって提案されている。

【０００５】

【発明が解決しようとする課題】しかしながら、上記のように宛て先をグループ化するなどしても、まだ大容量の記憶機構が必要となる局面が生じたり、宛て先情報の欠落をブロードキャストで補うことによる性能上の問題が生じたりして、その実現が現実的でなくなる場合があった。

【０００６】本発明は、分散共有メモリを実現する際に必要となるアクセス情報のマルチキャストの宛て先を記録するための記憶機構の容量の削減と不要なマルチキャストによる性能低下とのバランスを向上させ、低コストで高性能な情報処理システム及び情報処理装置及びその制御方法を提供することを目的とする。

【０００７】

【課題を解決するための手段】上記の目的を達成するための本発明の情報処理システムは以下の構成を備える。すなわち、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムであって、前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ手段と、前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメ

ンテナンス手段とを備える。

【０００８】また、上記の目的を達成するための本発明の情報処理装置は以下の構成を備える。すなわち、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置であって、前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ手段と、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、前記メモリ空間にアクセスが発生した場合、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段とを備える。

【０００９】また、上記の目的を達成するための本発明の情報処理システムの制御方法は以下の工程を備える。すなわち、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムの制御方法であって、前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ工程と、前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、前記保持工程によって保持されたキャッシュ情報に基づいて前記キャッシュ工程のキャッシュメンテナンスを行うメンテナンス工程とを備える。

【００１０】また、上記の目的を達成する本発明の情報処理装置の制御方法は、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置の制御方法であって、前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ工程と、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、前記メモリ空間にアクセスが発生した場合、前記保持工程によって保持されているキャッシュ情報に基づいて前記キャッシュ工程におけるキャッシュメンテナンスを行うメンテナンス工程とを備える。

【００１１】

【発明の実施の形態】以下、図面を参照して本発明の好適な一実施形態を詳細に説明する。

【００１２】＜第１の実施形態＞図１は、本実施形態の情報処理システムの採用する並列計算機システムの構成を示す図である。計算機は複数のノード（１００，２００，３００，４００，５００）からなり、各々のノードはコンセントレータ６００を通じて光ファイバによりネットワークを構成している。コンセントレータは、ノード間の通信を管理する。

【００１３】各ノードは通常の計算機であり、各々１個または複数個のＣＰＵを持つ。各々のノードはアドレス空間を共有している。図２は本実施形態によるメモリアドレス空間を説明する図である。図２に示されるように、一つのＣＰＵからみえるアドレス空間には他のノードのアドレス空間が含まれる。なお、ノード間通信は、自分のアドレス上に見えている相手のメモリに対し、直接にデータを書き込むことで行う。

【００１４】ノード（１００，２００，３００，４００，５００）は光通信装置（１４０，２４０，３４０，４４０，５４０）を通じ、アービトレーション回線（１６０，２６０，３６０，４６０，５６０）、データ回線（１７０，２７０，３７０，４７０，５７０）の２本の光ファイバを介してコンセントレータ６００に接続することによりネットワークを構成している。アービトレーション回線（１６０，２６０，３６０，４６０，５６０）は、ノード（１００，２００，３００，４００，５００）とコンセントレータ６００内の光バスアービタ６１０を接続する。アービトレーション回線により、各ノードは光バスアービタ６１０と相互に通信することができる。データ回線（１７０，２７０，３７０，４７０，５７０）は、ノードとコンセントレータ内のスターカプラ６２０を接続する。スターカプラ６２０は、一端に光を入力すると他端からその光が均等に出力される。従って、一つのノードがデータ光線に光を発すると、その光を他の全てのノードで受け取ることができる。この結果、一つのノードがデータ回線に発した光通信を、全てのノードで受け取ることが可能となる。

【００１５】ノードは、通信する波長を選択することにより、任意のノードと通信を行う。例えば、ノードＡ１００がノードＢ２００と通信を行う場合は以下の手順が実行される。まず、ノードＡ１００がアービトレーション回線１６０を通じて光バスアービタ６１０にノードＢ２００との通信要求を送る。そして、光バスアービタ６１０はノードＢ２００が他のノードと通信中でないことを確認すると、どのノードも使用していない波長αをノードＡ，Ｂ間（１００，２００間）の通信に割当てる。アービトレーション回線（１６０，２６０）を通じてノードＡ，Ｂ（１００，２００）に波長αを使用して通信を行うように指示が出されると、ノードＡ，Ｂ（１００，２００）が波長αを使用してデータ回線（１７０，２７０）を通じて通信を行う。

【００１６】分散共有メモリは、このようなネットワークの上で光通信により実現される。例としてノードＡ１００のプロセッサ１１０（以下、ＣＰＵ１１０）が、ノードＢ２００のメモリ２３０をリードする様子を図３に示す。なお、（Ａ１）～（Ａ７）に示す動作は一連のものとして実施される。

【００１７】図３において、

（Ａ１）ノードＡ１００のＣＰＵ１１０が、ノードＢの

メモリ２３０に対してリードアクセスを発行する。
（Ａ２）ノードＡ１００の光通信装置１４０は、メインバス上のバスアクセスを監視している。メインバス上に、そのノードＡのメモリ１３０以外のアドレスに対してアクセスが発行されると、光通信装置１４０はそのバスアクセスを検出する。
（Ａ３）光通信装置１４０は、アクセス要求パケットをコンセントレータ６００を通じて、ノードＢの光通信装置２４０に対して送る。
（Ａ４）ノードＢの光通信装置２４０はアクセス要求パケットを受け取ると、その依頼にしたがってノードＢのメモリ２３０に対してのアクセスを代行する。
（Ａ５）ノードＢのメモリ２３０がアクセスに応答する。
（Ａ６）ノードＢの光通信装置２４０は、アクセスが終わると、ノードａの光通信装置１４０に対してアクノリッジを返す。
（Ａ７）アクノリッジを受け取ったノードＡの光通信装置１４０は、ノードＢのメモリ（２３０）の代わりにノードＡのＣＰＵ１１０のアクセスに応答する。
【００１８】また本実施形態では、分散共有メモリ上で、ディレクトリ方式の一貫性保持方式を採用したキャッシュに類似したシステムを搭載している。ディレクトリ方式のキャッシュの詳細については、「共有記憶型並列システムの実際」（コロナ社刊）に示されている。
【００１９】通常のディレクトリ方式のキャッシュ一貫性保持機構では、メモリの各キャッシュアドレス毎に（正確にはキャッシュ単位毎であるが、以下、これをキャッシュアドレス毎ということにする）そのアドレスをキャッシングしているノードを記録する。例えばノードがＡ～Ｈまであるとすれば、各キャッシュアドレス毎に８ｂｉｔのディレクトリを待つ。ディレクトリの各ビットは、ノードＡ～Ｈに対応する。もしノードＡ，Ｂ，Ｆ，Ｈにそのアドレスがキャッシングされていれば、そのキャッシュアドレスに対応したディレクトリの値は２進数で１１０００１０１となる。
【００２０】本実施形態では上述の様なディレクトリ方式の代わりに、各キャッシュアドレス毎に１ｂｉｔのフラグを持たせる。以降このフラグをキャッシングフラグと呼ぶ。キャッシングフラグは、そのキャッシュアドレスが他のノードにキャッシングされているかどうかを示す。例えばノードＡのメモリ１３０内のアドレス１８０００００がノードＡ，Ｂ，Ｈにキャッシングされていれば、アドレス１８０００００に対応するキャッシングフラグはＯＮとなる。キャッシュアドレス１８０００００がどのノードにもキャッシングされていないか、もしくはノードＡ自身のみにキャッシングされている場合には、キャッシュアドレス１８０００００に対応するキャッシングフラグはＯＦＦになる。
【００２１】ノード内のＣＰＵ間でのキャッシュの一貫性保持は、ＣＰＵに内蔵されたＭＥＳＩプロトコル（Modified,Exclusive,Shared,Invalidの４状態によりキャッシュを管理するプロトコル）によるスヌープキャッシュで管理されている（この実施形態ではＭＥＳＩプロトコルを使用するが、これは本発明を限定するものではない）。
【００２２】さて、各ノード間のキャッシュの一貫性保持は、外付けされたキャッシュ管理装置（１５０，２５０，３５０，４５０，５５０）により管理される。キャッシュ管理装置は内部バス上でアクセスを検出すると、そのアクセスアドレスに対応したキャッシングフラグを参照し、もしキャッシングフラグがＯＮになっていたらバス上に発行されたアクセス（キャッシュメンテナンス情報）を他のノードに伝達する。
【００２３】実際のキャッシュメンテナンス情報のマルチキャストは、コンセントレータ６００内の光バスアービタ６１０により行われる。光バスアービタ６１０は、各ノードのキャッシュ管理装置からの要求によりアクセスを各ノードに伝達する。光バスアービタ６１０にはＩＣＣ６３０と呼ばれるアクセスマルチキャスト情報のキャッシュがある。ＩＣＣ６３０は最近使用されたアドレスについて、そのアドレスへのアクセスをどのノードに伝えるべきかを記録するキャッシュである。光バスアービタ６１０はアクセス要求を受け取ると、ＩＣＣ６３０を参照してそのアクセスを伝達すべきノードが記録されていないかを調べる。もし記録されていたら（以下、ＩＣＣヒットという）、光バスアービタ６１０はそれらのノードにアクセスをマルチキャストする。もし記録されていなければ（以下、ＩＣＣミスという）、全てのノードにアクセスをブロードキャストする。
【００２４】以下、キャッシュ管理装置及び光バスアービタの構成について説明する。
【００２５】図４はノードＡにおけるキャッシュ管理装置１５０の構成を示すブロック図である。ノード間のキャッシュの一貫性保持をするキャッシュ管理装置１５０は、バス監視装置１５１、キャッシングフラグ１５２、キャッシュフラグ管理装置１５３、通信管理装置１５４、アクセスエミュレーション装置１５５を備える。以下にキャッシュ管理装置１５０の備える各装置について説明する。
【００２６】図５は、バス監視装置１５１の構成を示すブロック図である。バス監視装置１５１の内部は、シーケンサ１５１ａとアドレスラッチ１５１ｂからなっている。シーケンサ１５１ａはノード内のバスを監視し、バスマスタ（例えばＣＰＵ１１０）によるメモリ１３０のリードアクセス、invalitate,Read-with-Intent-to-Modify、メモリ書込みの実行を検出する。シーケンサ１５１ａは、バス上のアクセス要求信号、バスアクセスに伴ってバスマスタから出力されるアクセス修飾信号などを監視することで、アクセスを検出する。

【００２７】バス監視装置１５１はアクセス要求信号が有効であり、アクセスアドレスが自分のノードのメモリに対するものであったとき、もしくは他ノードのメモリへのキャッシュメンテナンスアクセスであったとき、アクセスを検出したと判断する。アクセスが自ノードのメモリに対するものであったとき、バス監視装置１５１はアクセスを検出するとメモリ１３０の応答を押さえ、その間にキャッシュフラグ管理装置１５３を通じてキャッシングフラグ１５２を調べる。もし対応するキャッシングフラグがＯＮなら、バス監視装置１５１は光バスアービタ６１０にアクセスの伝達要求を出す。そしてそれらに対する応答により対象アドレスが他のノードにキャッシングされているかどうかを知り、キャッシュフラグ管理装置１５３を通じてキャッシュフラグ１５２を更新する。その後バス監視装置１５１がメモリ１３０の応答を許可することにより、ＣＰＵ１１０はそのアクセスを完了する。一方、他ノードのメモリへのキャッシュメンテナンスアクセスであったときは、ＣＰＵ１１０のアクセスをリトライさせておき、その間に光バスアービタ６１０にアクセスの伝達要求を出す。そして、その完了パケットが届いたらＣＰＵ１１０のアクセスリトライを解除し、処理を続行させる。

【００２８】なお、上記メモリアクセスの中断は、バス上にアクセス再実行信号を出力し、そのメモリアクセスをプロセッサにリトライさせることで実現したが、他の方法によりアクセスを中断してもよい。

【００２９】図６はキャッシュフラグ１５２及びキャッシュフラグ管理装置の構成を説明するブロック図である。先に述べたように本実施形態は、メモリの各キャッシュブロックに対して１ｂｉｔのキャッシュフラグを持つ。キャッシングフラグは、そのアドレスが他のノードにキャッシングされているかどうかを示す。キャッシュフラグ１５２への操作は、キャッシュフラグ管理装置１５３により行われる。キャッシュフラグ管理装置１５３はシーケンサ１５３ａにより管理されている。キャッシュフラグ１５２は、バス監視装置１５１が他ノードからのリード／ライトアクセスを検出した際に、対象メモリブロックが外部ノードにキャッシングされていることを記録する場合にＯＮにされ、バス管理装置１５１がメモリ書込み／invalidata／Read-with-Intent-to-Modifyを検出した場合、或いは対象メモリブロックが自ノードにのみキャッシングされている場合にはＯＦＦにされる。

【００３０】通信管理装置１５４はバス監視装置１５１、キャッシュフラグ管理装置１５３からの要求により、光通信装置１４０とのコミュニケーションを行う。通信管理装置１５４は、バス監視装置１５１、キャッシュフラグ管理装置１５３からのリード／invalidata／Read-with-Intent-to-Modify／メモリ書込みのマルチキャスト要求を受け、光通信装置１４０を介して光バスアービタ６１０に、アクセスのマルチキャスト／ブロードキャストパケットを送出する。また光通信装置１４０を通じて受け取ったキャッシュの一貫性保持動作要求パケットにより、メモリアクセスエミュレート装置１５５に対してinvalidata／Read-with-Intent-to-Modify／ライト／リード要求を出力する。

【００３１】図７はメモリアクセスエミュレート装置の構成を示すブロック図である。メモリアクセスエミュレート装置１５５は、通信管理装置１５４からの要求により、ダミーのメモリライトアクセス／ダミーのメモリリードアクセス／invalidata／Read-with-Intent-to-Modifyアクセスを自ノードのバス上に発行する。これらのアクセスにより、ノード間のキャッシュの一貫性を保持する。エミュレート装置１５５の発行するアクセスは、自分のノードに割り当てられたアドレスではありえないため、ノード内のバススレーブ（例えばメモリ１３０）より応答はありえず、ダミーのアクセスとなる。

【００３２】次に、コンセントレータ６００について説明する。図８はコンセントレータ６００の構成を示すブロック図である。コンセントレータ６００は、光バスアービタ６１０、スターカプラ６２０、ＩＣＣ６３０を備える。光バスアービタ６１０は先に述べたように光回線の通信制御を行う。またノードから送られてきたマルチキャスト要求により、各ノードへのパケットの転送を行う。また、ＩＣＣ６３０はアクセスのマルチキャスト先を保持するキャッシュメモリである。光バスアービタ６１０は、マルチキャスト対象アドレスがＩＣＣ６３０にヒット（ＩＣＣヒット）した場合にはＩＣＣ６３０に記録されたノードに対してマルチキャストを行う。ＩＣＣ６３０にミス（ＩＣＣミス）した場合は全てのノードに対してブロードキャストを行う。

【００３３】ＩＣＣ６３０を用いることにより、マルチキャストの必要なノードを正確に知ることができるようになり、不必要なブロードキャストがなくなる。このため、当該情報処理システムの性能を向上することができる。図９はＩＣＣ６３０のデータ構成を示す図である。本実施形態ではＩＣＣ６３０は図９のようなフルマップ（各ノードにつき１ビットを割り当てる）方式をとっている。

【００３４】なお、本実施形態では、光バスアービタ６１０は一つの計算機として実現するので、これらの機構はソフトウェアで実現されるものとする。しかしながら、この構成例は本発明を制限するものではなく、ハードウェアで実現することも可能であることは明らかである。

【００３５】以上の様な構成を備える本実施形態の動作について以下に詳細に説明する。

【００３６】図１０はキャッシュ管理装置の動作を説明するフローチャートである。まず、ステップＳ１０１において、バス監視装置１５１がバスアクセスを検出するとステップＳ１０２へ進む。ステップＳ１０２では、当

該アクセスが自ノード内のメモリに対するアクセスか否かを判定する。そして、自ノード内のメモリへのアクセスであった場合はステップＳ１０３へ、そうでない場合はステップＳ１１１へそれぞれ進む。

【００３７】ステップＳ１０３では、バス監視装置１５１がメモリ１３０の応答を抑制する。そして、キャッシュフラグ管理装置１５３を通してキャッシュフラグ１５２を参照し、当該キャッシュアドレスが他ノードにキャッシングされているか否かを判定する。他ノードへにキャッシングされていなければ、そのままステップＳ１０８へ進み、メモリ１３０に対する応答抑制を解除し、当該アクセスを完了させる。

【００３８】一方、ステップＳ１０４で当該キャッシュアドレスが他ノードに対してキャッシングされていると判定された場合は、他のノードに当該アクセスを通知するためにステップＳ１０５へ進む。ステップＳ１０５では、通信管理装置１５４に対して当該アクセスを通知するべく指示を出す。通信管理装置１５４は光通信装置１４０を通じて、コンセントレータ６００にマルチキャスト要求を通知する。そして、コンセントレータ６００よりマルチキャスト完了パケット（マルチキャスト完了パケットには、当該キャッシュアドレスが他ノードへキャッシングされている否かを示す情報が含まれる）を受信すると、このパケットに従ってキャッシュフラグ１５２を更新する。

【００３９】その後、ステップＳ１０８にて、メモリ１３０に対する応答抑制を解除し、当該アクセスを完了させる。また、ステップＳ１０２で自ノードのメモリではないと判定された場合は、ステップＳ１１１へ進み、当該アクセスがキャッシュメンテナンスを含むか否かを判定する。キャッシュメンテナンス情報を含まない場合は、図３で説明した様な手順によって、光通信装置からアクセス要求パケットが発行されることになる。また、キャッシュメンテナンスを含むアクセスであれば、ステップＳ１１２へ進み、メモリをリトライさせる。そして、ステップＳ１１３において、対象メモリを有するノードに対してアクセス要求パケットを送出し、その応答を待つ。そして、応答を得たならば、ステップＳ１１４にてメモリのリトライを解除し、本アクセスを完了する。なお、ステップＳ１１１～Ｓ１１４に関しても、キャッシュ管理装置１５０内のバス管理装置１５１、通信管理装置１５４が処理している。

【００４０】図１１は、光バスアービタよりパケットを受信した場合の動作を説明するフローチャートである。光バスアービタ６１０よりパケットを受信すると、アクセスエミュレーション装置１５５は当該パケットに従ってバス上にアクセス情報を出力する（ステップＳ１５１、Ｓ１５２）。この結果、バスをスヌーブしているＣＰＵ１１０、１２０やバス監視装置１５１が応答情報を生成し、通信管理装置１５４がこの応答情報をアクセス結果として光バスアービタ６１０に通知する（ステップＳ１５３）。

【００４１】図１２は、光バスアービタの動作手順を説明するフローチャートである。図１２では、キャッシュメンテナンスに係るマルチキャスト（ブロードキャストを含む）に対する応答動作が示されている。

【００４２】ステップＳ２０１において、キャッシュアドレスを含むマルチキャスト要求パケットを受信すると、ステップＳ２０２へ進み、当該キャッシュアドレスでＩＣＣ６３０を検索する。ＩＣＣミスした場合は、ステップＳ２０３へ進み、全ノードに対してマルチキャスト（ブロードキャスト）を行い、ステップＳ２０５へ進む。一方、ステップＳ２０２においてＩＣＣヒットした場合は、ステップＳ２０４へ進み、当該キャッシュアドレスに対応して記録されているノードに対してマルチキャストを行い、ステップＳ２０５へ進む。

【００４３】ステップＳ２０５では、マルチキャスト先の各ノードからの完了パケットを待つ。各ノードからの完了パケットを受信したら、ステップＳ２０６へ進み、受信した完了パケットに従ってＩＣＣ６３０を更新する。そして、ステップＳ２０７において、当該マルチキャスト要求もとのノードに対してマルチキャスト完了パケットを送出する。

【００４４】次にキャッシュ管理装置１５０の動作を更に具体的に説明する。すなわち、

・ノード内のアドレスにリード、invalidate、Read-with-Intent-to-Modify（キャッシュのライトミス時に出力されるライトを前提としたリードサイクルであり、以下ＲＷＩＴＭとする）などのキャッシュメンテナンス情報が出力されたとき、及び

・ノード外のアドレスに対してinvalidate、ＲＷＩＴＭなどのキャッシュメンテナンス情報が出力されたときについて具体的な動作を説明する。

【００４５】図１３は、ノード内のメモリに対してリード／ライト／invalidate／ＲＷＩＴＭ等のバスアクセスが発行された場合の動作を説明する図である。

（Ｂ１）ノード内のメモリ１３０に対してリードアクセスが行われる。

（Ｂ２）バス監視装置１５１がそのアクセスを検知する。

（Ｂ３）メモリ１３０へのアクセスを検知したバス監視装置１５１は、メモリ１３０の応答を抑制する。

（Ｂ４）その間にバス監視装置１５１はキャッシュフラグ管理装置１５３に、当該キャッシュアドレスに関する参照要求を出す。

（Ｂ５）キャッシュフラグ管理装置１５３はキャッシュフラグ１５２に指定されたキャッシュアドレスが他ノードにキャッシングされているか（対象アドレスに対応したビットがＯＮかＯＦＦか）調べる。

（Ｂ６）もし対応するキャッシュフラグがＯＮならば、

キャッシュフラグ管理装置１５３は通信管理装置１５４を通じて発行されたリードアクセスを他のノードにマルチキャストするよう光バスアービタ６１０に要求を出す。

（Ｂ７）光バスアービタ６１０はＩＣＣ６３０を参照する。

（Ｂ８）ＩＣＣミスならば光バスアービタ６１０は全てのノードに当該リードアクセスをブロードキャストする。

（Ｂ９）リードアクセスをブロードキャストされた各ノード（２００，３００，４００，５００）は、バス上にリードアクセスを発行する。

（Ｂ１０）各ノードのＣＰＵ（２１０，３１０，４１０，５１０）はこれをスヌープし、そのリードアクセスされたメモリ番地をキャッシングしているかどうかを返答する。

（Ｂ１１）各ノードのバス監視装置（２５１，３５１，４５１，５５１）はこれを検出し、通信管理装置（２５４，３５４，４５４，５５４）を通じて、アクセスが完了したこととそのノードで該当アドレスがキャッシングしているかどうかを光バスアービタ６１０に伝える。

（Ｂ１２）各ノードからリードアクセスが完了したことを通知された光バスアービタ６１０は、各ノードがキャッシングしている／していないという情報をＩＣＣ６３０にキャッシングし、ＩＣＣ６３０を更新する。

（Ｂ１３）他ノードに対象番地がキャッシングされているかどうかの情報を含んだアクセス完了パケットをノードＡ１００に伝える。

（Ｂ１４）通信管理装置１５４を通じて各ノードのキャッシング状況を受け取ったキャッシュフラグ管理装置１５３は、その内容をキャッシュフラグ１５２に記録する。

（Ｂ１５）記録が終了すると、バス監視装置１５１はメモリ１３０の応答抑制を解除し、当該リードアクセスが完了する。

【００４６】もし（Ｂ８）においてＩＣＣヒットならば、光バスアービタ６１０は記録されたノードにアクセスをマルチキャストする。その後の動作は上記の物と同様である。

【００４７】上記手順によれば、最初の１回目のアクセスは、アクセス情報を必要のないノードに対してもブロードキャストを生じさせる。しかし、必要のないノードに対してのブロードキャストは、システムの性能を落とすことにはなるが論理的な矛盾を生じさせるものではない。各ノードがブロードキャストに応答して返してくるパケットの中に、そのノードが該当アドレスをキャッシングしているかの情報が含まれているので、それに基づきキャッシュフラグをアップデートすることで、２回目以降のアクセスは必要なノードにのみマルチキャストが行われることになる。

【００４８】また、ノード外のアドレスに対してリード／ライト／invalidata等のバスアクセスが出力された時は、光通信装置を通じて他ノードにアクセスを依頼する。アクセス先のノードのキャッシュ管理装置は、先に述べたように光通信装置を通じて行われたアクセスに対して応答する。

【００４９】以上の様に、上記実施形態によれば、すべてのキャッシュアドレスについてそのアドレスが他の計算機にキャッシングされているかどうかを示すキャッシングフラグが記録され、他のノードにキャッシングされているアドレスに対してのメモリアクセスのみが他の計算機に伝達されることになる。このため、効率のよいマルチキャスト要求が実現されることになる。

【００５０】なお、以上のような構成をとるのは、計算機外への伝達コストが比較的高いためである。計算機外に伝達が必要なアクセスが多発した場合には、システム全体としての性能を発揮することができない。そこで外部へのアクセスであるかどうを的確に判定し、計算機外への不要なアクセスを無くす或いは低減することでシステムの性能の向上が図られる。

【００５１】＜第２の実施形態＞次に本発明の第２の実施形態を、図面を参照して説明する。

【００５２】第１の実施形態ではＩＣＣ６３０の構造を、フルマップ構造（ひとつのノードを１ビットで示す方法）で記録するようになっていた。このような場合、システムのノード数が多くなればＩＣＣ６３０が必要とするメモリ容量も増加する。、また、メモリ容量の増大化を防止すれば、その分登録可能なキャッシングアドレスが減少してしまう。そこで第２の実施形態はＩＣＣ６３０の構成を変更し、１つのキャッシングアドレスに関する記憶容量の削減を図る。

【００５３】図１４は、第２の実施形態におけるＩＣＣ６３０の構造を示す図である。ＩＣＣ６３０の各データレコードは、キャッシュタグ６３０ａ、ノードグループ６３０ｂ、マップ情報６３０ｃで構成される。キャッシュタグ６３０ａはキャッシングアドレスを示す。ノードグループ６３０ｂには、ノードグループ内の先頭のノードが格納される。本例では、ノードはＡからＺまでの順に順序付けがなされており、ノードグループ６３０ｂによって先頭のノードとして指定されたノードから順に６つのノードがノードグループとなる。例えば、ノードグループ６３０ｂにノードＡが指定されていれば、ノードＡからノードＦがノードグループとして指定されたことになる。

【００５４】ＩＣＣ６３０の下位はマップ情報６３０ｃであり、これは、当該グループ内のノードについてのフルマップの情報である。例えば上位でノードＳが示されていた場合、下位ではノードＳ～Ｘについてのフルマップの情報が示される。同様に上位でノードＣが示されていた場合、下位ではノードＣ～Ｈがフルマップで示され

ている。ノードＡとＪが同時に同じ番地をキャッシングした場合はこのような形式では対応できないので、ＩＣＣ６３０の上位を特別のパターン（例えば全て１）にして、その番地に対するアクセスはブロードキャスト（全てのノードに対して放送する）する。

【００５５】以上の様に第２の実施形態によれば、グループ内で局所的なメモリ共有が行われている場合に、このような構成にすることでＩＣＣ６３０の容量の削減を図ることができる。

【００５６】以上説明したように、上記各実施形態によれば、各ノードに大容量のディレクトリ（メモリ）を備えることなく、フルマップ法式を採用したディレクトリ方式に匹敵する高性能のキャッシュシステムを構築することができる。

【００５７】なお、本発明は、複数の機器（例えばホストコンピュータ，インタフェイス機器，リーダ，プリンタなど）から構成されるシステムに適用しても、一つの機器からなる装置（例えば、複写機，ファクシミリ装置など）に適用してもよい。

【００５８】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読出し実行することによっても、達成されることは言うまでもない。

【００５９】この場合、記憶媒体から読出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【００６０】プログラムコードを供給するための記憶媒体としては、例えば、フロッピディスク，ハードディスク，光ディスク，光磁気ディスク，ＣＤ－ＲＯＭ，ＣＤ－Ｒ，磁気テープ，不揮発性のメモリカード，ＲＯＭなどを用いることができる。

【００６１】また、コンピュータが読出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００６２】さらに、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００６３】

【発明の効果】以上説明したように、本発明によれば、分散共有メモリを実現する際に必要となるアクセス情報のマルチキャストの宛て先を記録するための記憶機構の容量の削減と不要なマルチキャストによる性能低下とのバランスを向上させ、分散共有メモリシステムにおいて低コストで高性能なキャッシュシステムを提供できる。

【００６４】

【図面の簡単な説明】

【図１】本実施形態の情報処理システムの採用する並列計算機システムの構成を示す図である。

【図２】本実施形態によるメモリアドレス空間を説明する図である。

【図３】ノードＡ１００のＣＰＵ１１０が、ノードＢ２００のメモリ２３０をリードする様子を説明する図である。

【図４】ノードＡにおけるキャッシュ管理装置１５０の構成を示すブロック図である。

【図５】バス監視装置１５１の構成を示すブロック図である。

【図６】キャッシュフラグ１５２及びキャッシュフラグ管理装置の構成を説明するブロック図である。

【図７】メモリアクセスエミュレート装置の構成を示すブロック図である。

【図８】コンセントレータ６００の構成を示すブロック図である。

【図９】ＩＣＣ６３０のデータ構成を示す図である。

【図１０】キャッシュ管理装置の動作を説明するフローチャートである。

【図１１】光バスアービタよりパケットを受信した場合の動作を説明するフローチャートである。

【図１２】光バスアービタの動作手順を説明するフローチャートである。

【図１３】ノード内のメモリに対してリード／ライト／invalidate／ＲＷＩＴＭ等のバスアクセスが発行された場合の動作を説明する図である。

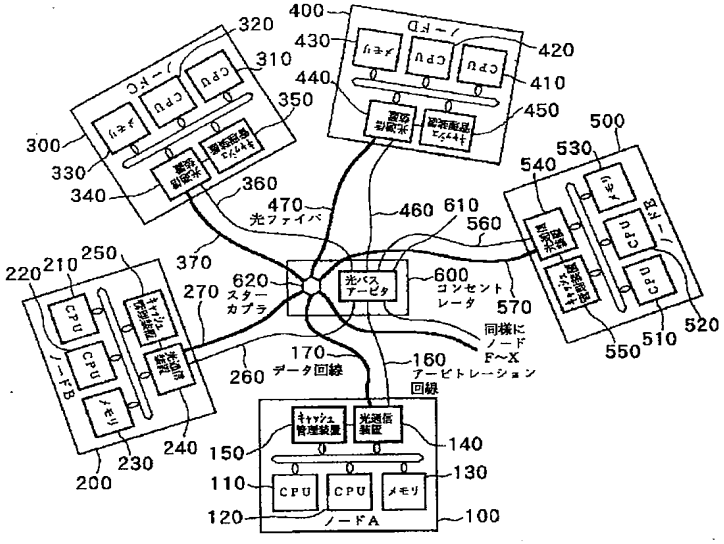
【図１４】第２の実施形態におけるＩＣＣ６３０の構造を示す図である。

【符号の説明】

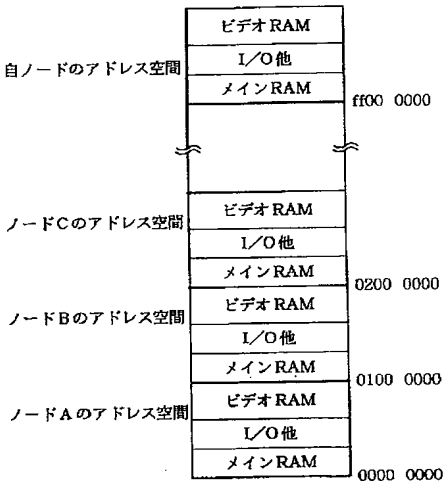
１００，２００，３００，４００，５００　ノード
１１０，１２０，２１０，２２０，３１０，３２０，４１０，４２０，５１０，５２０　ＣＰＵ
１３０，２３０，３３０，４３０，５３０　メモリ
１４０，２４０，３４０，４４０，５４０　光通信装置
１５０，２５０，３５０，４５０，５５０　キャッシュ管理装置
１６０，２６０，３６０，４６０，５６０　アービトレーション回線
１７０，２７０，３７０，４７０，５７０　データ回線
６００　コンセントレータ

６１０　光バスアービタ　　＊　＊６２０　スターカプラ

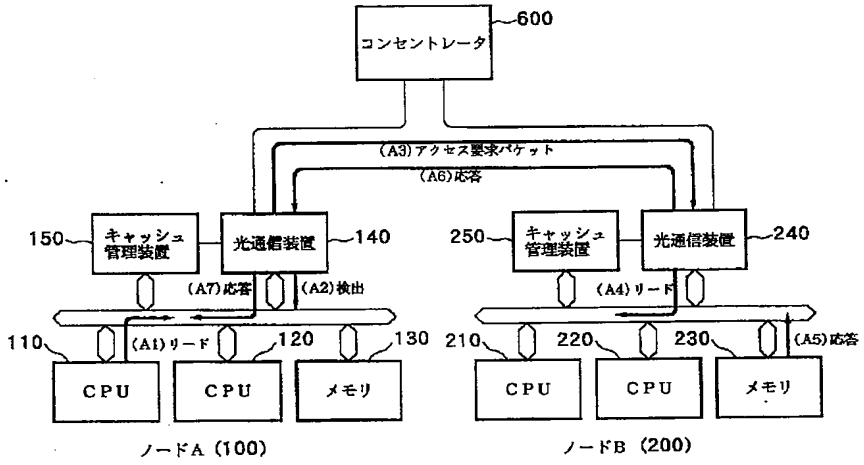
【図１】



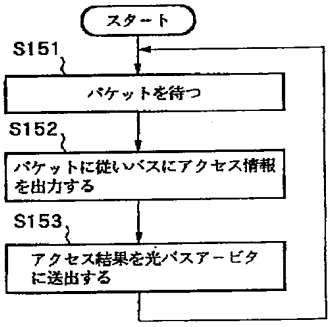
【図２】



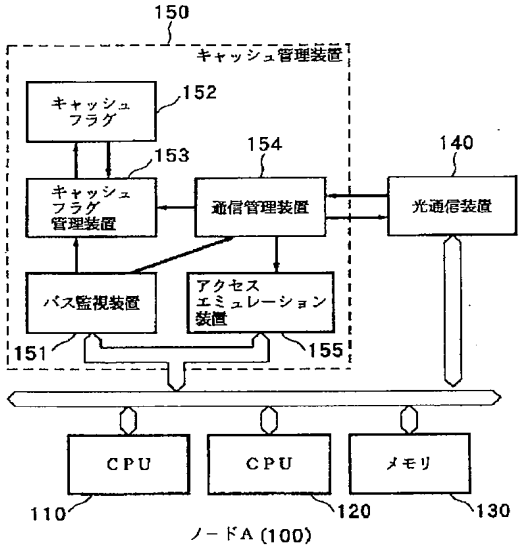
【図３】



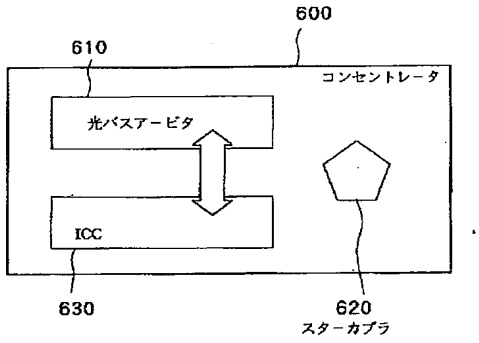
【図１１】



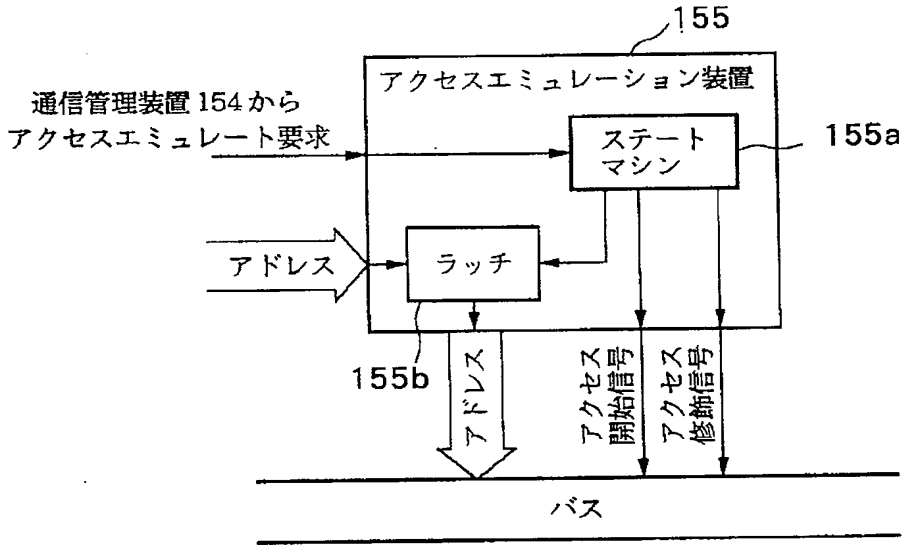
【図４】


150
キャッシュ管理装置
キャッシュフラグ
152
153
154
140
キャッシュフラグ管理装置
通信管理装置
光通信装置
バス監視装置
アクセスエミュレーション装置
151
155
CPU
CPU
メモリ
110
120
130
ノードA (100)

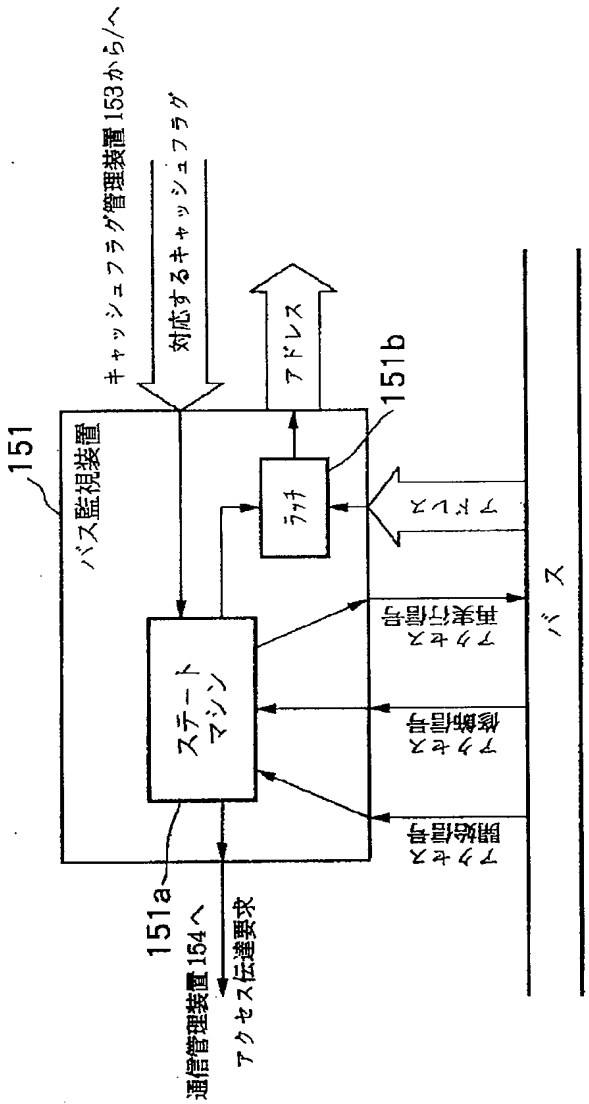
【図８】


600
610
コンセントレータ
光バスアービタ
ICC
630
620
スターカプラ

【図７】


155
アクセスエミュレーション装置
通信管理装置154から
アクセスエミュレート要求
ステートマシン
155a
アドレス
ラッチ
155b
アドレス
アクセス開始信号
アクセス修飾信号
バス

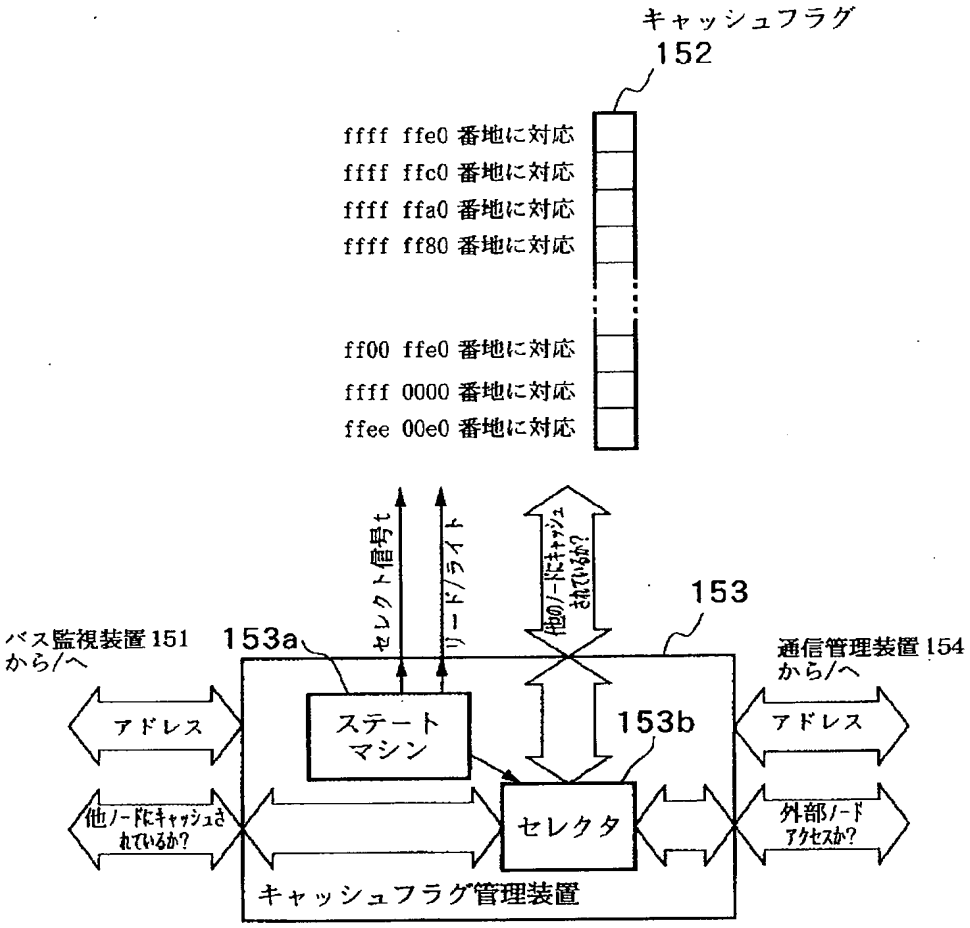
【図５】


151
バス監視装置
キャッシュフラグ管理装置153から/へ
対応するキャッシュフラグ
アドレス
151b
ラッチ
アドレス
ステートマシン
アクセス再実行信号
アクセス修飾信号
アクセス開始信号
バス
151a
通信管理装置154へ
アクセス伝達要求

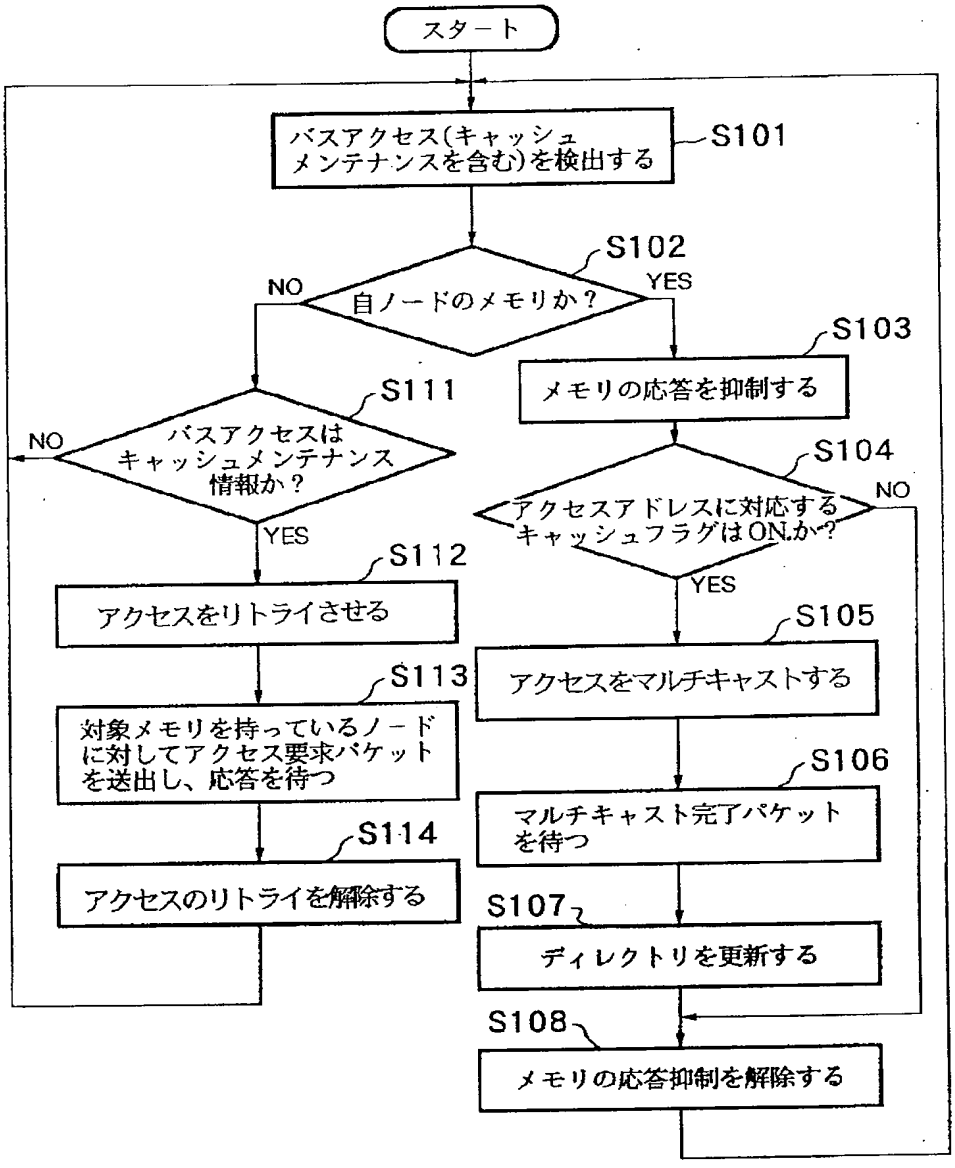
【図９】

ノードAがキャッシングしているか
ノードBがキャッシングしているか
ノードCがキャッシングしているか
ノードVがキャッシングしているか
ノードWがキャッシングしているか
ノードXがキャッシングしているか
キャッシュタグ
ffff ffe0
ffff ffc0
ffff ffa0
ffff ff80
ff00 ffe0
ffff 0000
ffee 00e0
ffff ffe0番地のディレクトリ情報
ffff ffc0番地のディレクトリ情報
ffff ffa0番地のディレクトリ情報
ffff ff80番地のディレクトリ情報
ff00 ffe0番地のディレクトリ情報
ffff 0000番地のディレクトリ情報
ffee 00e0番地のディレクトリ情報

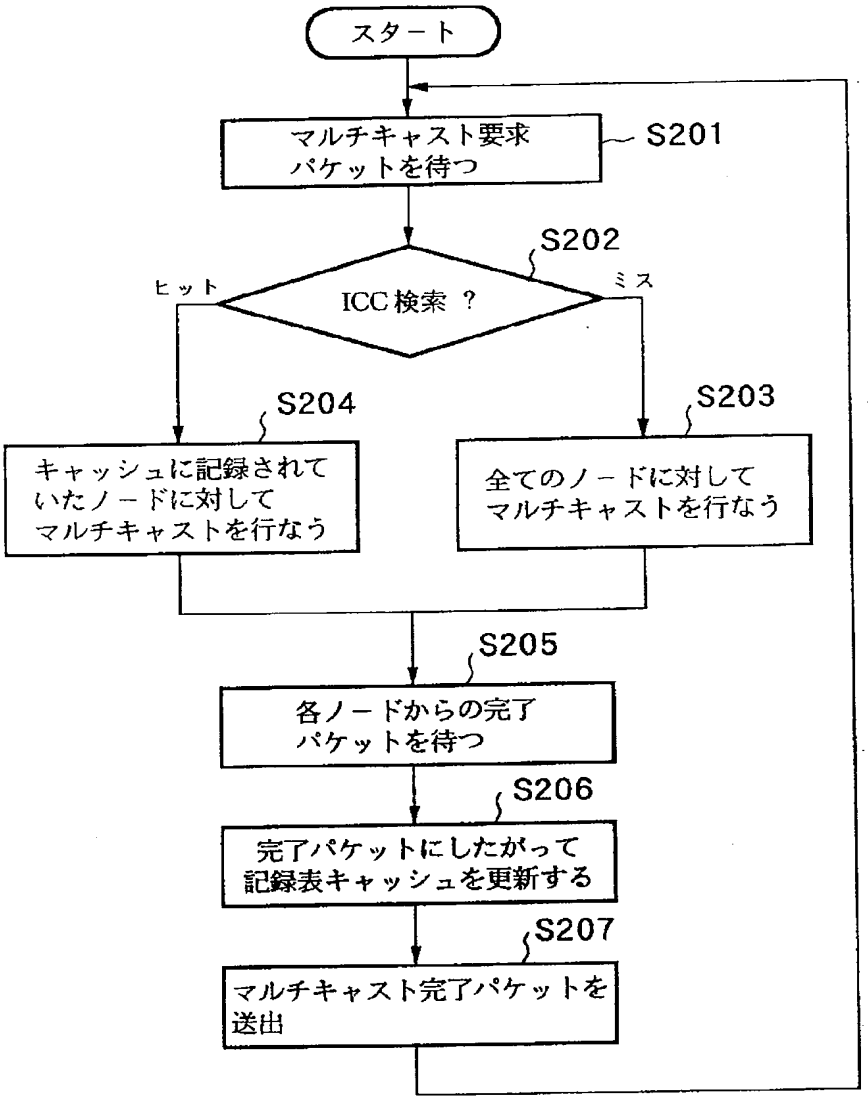
【図６】


キャッシュフラグ
152
ffff ffe0 番地に対応
ffff ffc0 番地に対応
ffff ffa0 番地に対応
ffff ff80 番地に対応
ff00 ffe0 番地に対応
ffff 0000 番地に対応
ffee 00e0 番地に対応
セレクト信号t
リード/ライト
他のノードにキャッシュされているか?
153
バス監視装置 151
から/へ
153a
通信管理装置 154
から/へ
アドレス
ステート
マシン
153b
アドレス
他ノードにキャッシュされているか?
セレクタ
外部ノードアクセスか?
キャッシュフラグ管理装置

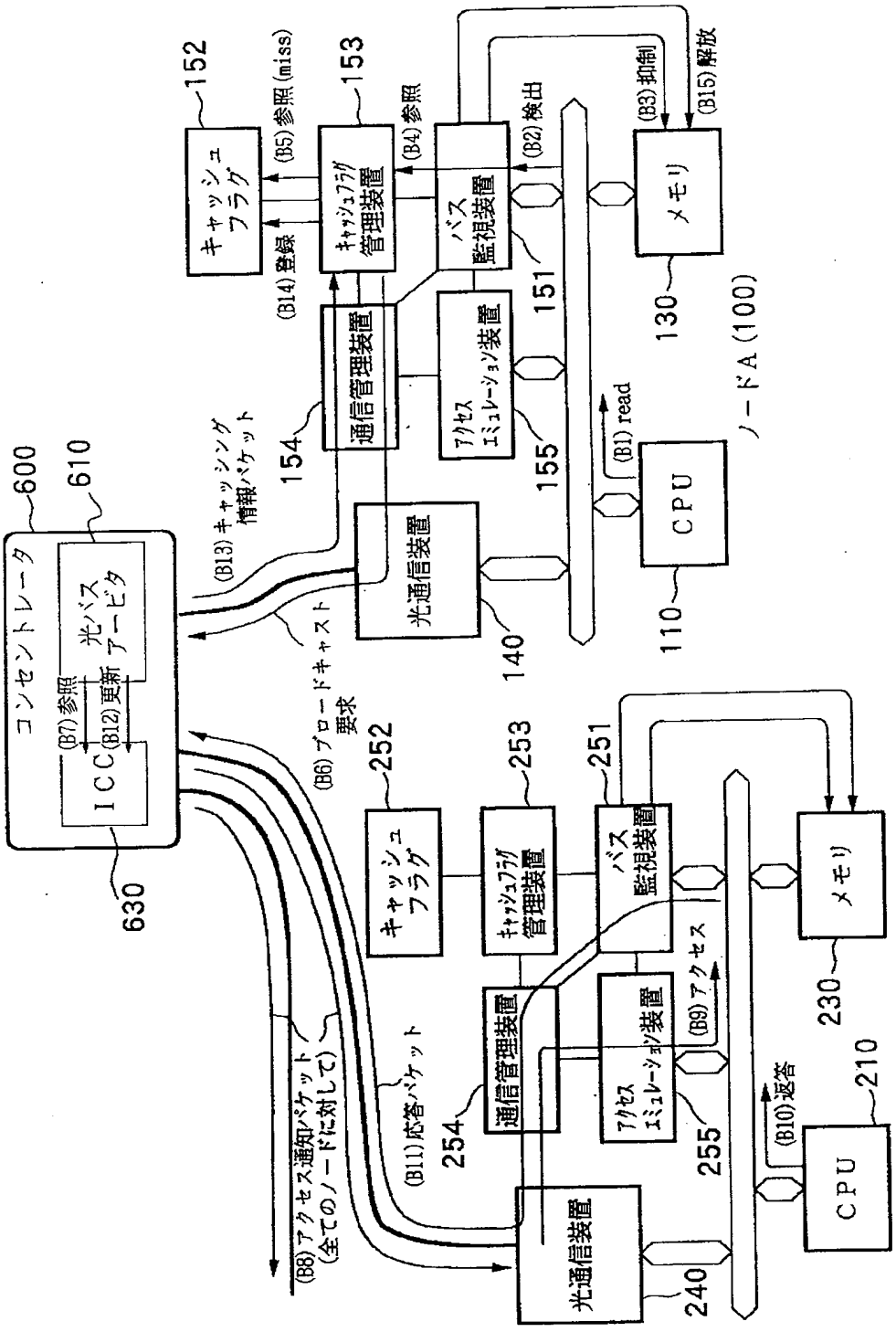
【図１０】


スタート
バスアクセス(キャッシュメンテナンスを含む)を検出する
S101
S102
自ノードのメモリか？
NO
YES
S103
メモリの応答を抑制する
S104
アクセスアドレスに対応するキャッシュフラグはON.か?
NO
YES
S105
アクセスをマルチキャストする
S106
マルチキャスト完了パケットを待つ
S107
ディレクトリを更新する
S108
メモリの応答抑制を解除する
S111
バスアクセスはキャッシュメンテナンス情報か？
NO
YES
S112
アクセスをリトライさせる
S113
対象メモリを持っているノードに対してアクセス要求パケットを送出し、応答を待つ
S114
アクセスのリトライを解除する

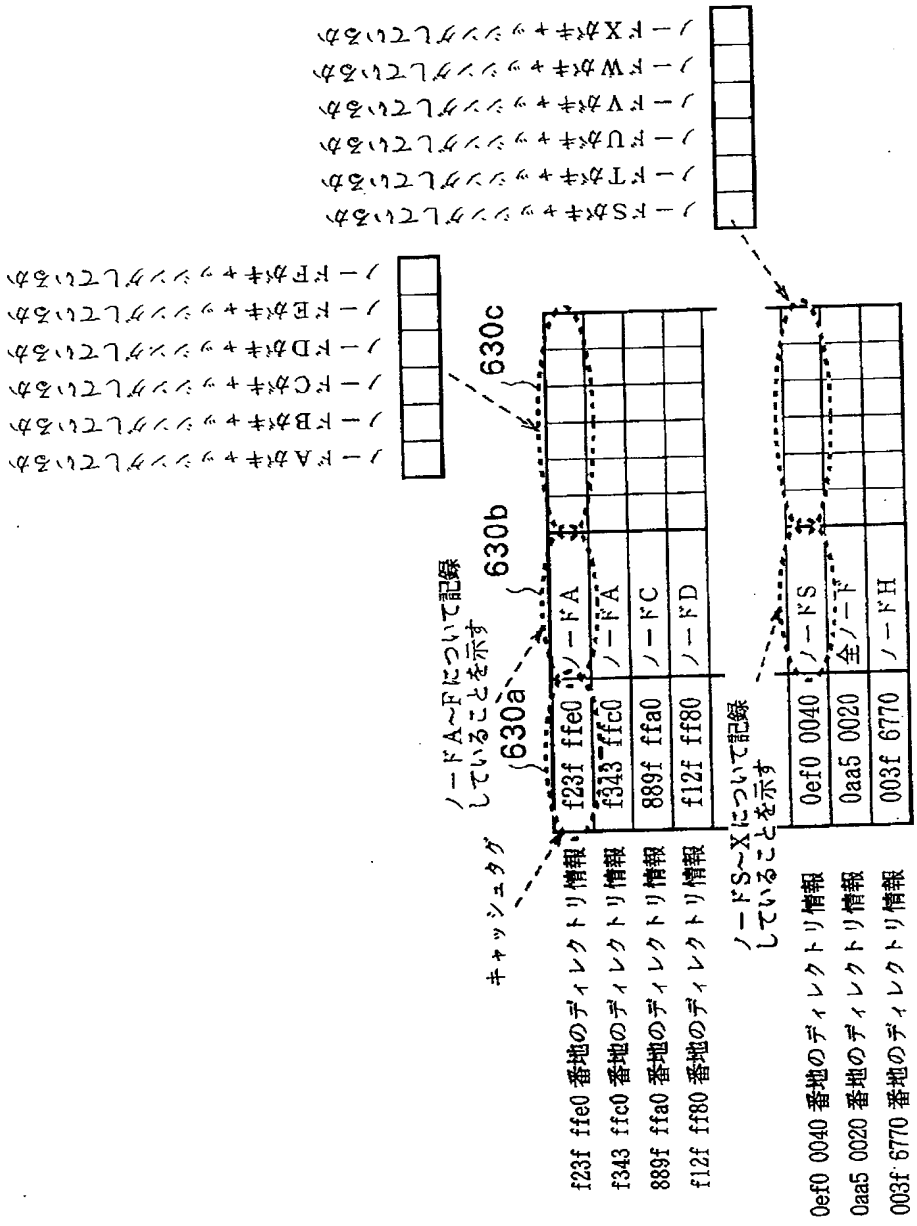
【図１２】


スタート
マルチキャスト要求
パケットを待つ
S201
S202
ICC 検索 ？
ヒット
ミス
S204
キャッシュに記録されて
いたノードに対して
マルチキャストを行なう
S203
全てのノードに対して
マルチキャストを行なう
S205
各ノードからの完了
パケットを待つ
S206
完了パケットにしたがって
記録表キャッシュを更新する
S207
マルチキャスト完了パケットを
送出

【図１３】


コンセントレータ
600
610
光バス
アービタ
ICC
(B7)参照
(B12)更新
630
(B13)キャッシング
情報パケット
(B6)ブロードキャスト
要求
(B8)アクセス通知パケット
(全てのノードに対して)
(B11)応答パケット
154
通信管理装置
光通信装置
140
キャッシュフラグ
152
(B14)登録
(B5)参照(miss)
キャッシュフラグ
管理装置
153
(B4)参照
バス
監視装置
(B2)検出
151
アクセス
エミュレーション装置
155
(B1) read
CPU
110
(B3)抑制
(B15)解放
メモリ
130
ノードA(100)
252
253
251
キャッシュフラグ
キャッシュフラグ
管理装置
バス
監視装置
メモリ
230
通信管理装置
254
アクセス
エミュレーション装置
255
(B9)アクセス
(B10)返答
光通信装置
240
CPU
210

【図１４】


ノードA～Fについて記録
していることを示す
630a
630b
630c
キャッシュタグ
f23f ffe0
f343 ffc0
889f ffa0
f12f ff80
ノードA
ノードA
ノードC
ノードD
f23f ffe0 番地のディレクトリ情報
f343 ffc0 番地のディレクトリ情報
889f ffa0 番地のディレクトリ情報
f12f ff80 番地のディレクトリ情報
ノードS～Xについて記録
していることを示す
0ef0 0040
0aa5 0020
003f 6770
ノードS
全ノード
ノードH
0ef0 0040 番地のディレクトリ情報
0aa5 0020 番地のディレクトリ情報
003f 6770 番地のディレクトリ情報